UNITED STATES PATENT APPLICATION FOR:

METHOD AND APPARATUS FOR PERFORMING RELATIONAL SPEECH RECOGNITION

INVENTORS:

HORACIO E. FRANCO
DAVID J. ISRAEL
GREGORY K. MYERS

ATTORNEY DOCKET NUMBER: SRI/4580-2

CERTIFICATION OF MAILING UNDER 37 C.F.R. 1.10

I hereby certify that this New Application and the documents referred to as enclosed therein are being deposited with the United States Postal Service on Sept. 28, 2001, in an envelope marked as "Express Mail United States Postal Service", Mailing Label No. EL728297870 US, addressed to: Assistant Commissioner for Patents, Box PATENT APPLICATION, Washington, D.C. 20231.

Kathleen Faughnan
Signature

Kathleen Faughnan
Name

9-28-01
Date of signature

MOSER, PATTERSON & SHERIDAN, LLP
595 Shrewsbury Ave.
Shrewsbury, New Jersey 07702
(732)530-9404

METHOD AND APPARATUS FOR PERFORMING RELATIONAL SPEECH RECOGNITION

BACKGROUND OF THE INVENTION

Field of the Invention

[0001] The invention relates generally to speech recognition and, more specifically, to speech recognition systems used to recognize groups of words that have observable relationships.

Description of the Related Art

[0002] Global Positioning System (GPS)-based navigation systems have recently become available in some automobiles. To use these systems, a driver must enter an address or location, typically with a touch screen. The navigation system then provides instructions (usually with a map, arrow displays, or a synthesized voice) that directs the driver from the present location to the desired location.

[0003] Although current navigation systems work quite well at providing directions, a driver cannot enter new locations via the touch screen while the car is moving. And even when the car is stopped, using a touch screen to enter an address can be slow and difficult.

[0004] Replacing or supplementing the touch screen with a speech recognition system would make navigation systems much easier to use, and would make it possible for a driver to enter an address while the car is moving. However, it is well known that recognition of spoken addresses is an extremely difficult task because of the huge number of street and city names that such a speech recognition system would need to recognize. See, for example, the discussion in U.S. Patent No. 5,177,685 to Davis et al., at column 24, lines 45-56.

[0005] One way to reduce the "search space" of a speech recognizer is to use a "prompt and response" type interface. These systems typically prompt the speaker to say only one word or short phrase at a time. For example, the speaker may be prompted to say only a street number or only a city name. This allows the system to perform a series of much simpler speech recognition passes, rather than performing the very difficult task of recognizing an entire spoken address or other long phrase.

[0006] Although prompt and response systems simplify the speech recognition task, they can be both slow and annoying to use because they require the speaker to answer a series of questions. Accordingly, there remains a need for speech recognition system that allows the speaker to say an address or other difficult to recognize phrase in a single utterance and have that utterance understood and acted upon.

SUMMARY OF THE INVENTION

[0007] The disadvantages associated with the prior art are overcome by a method and apparatus for performing speech recognition using the observable relationships between words. Results from a speech recognition pass can be combined with information about the observable word relationships to constrain or simplify subsequent recognition passes. This iterative process greatly reduces the search space required for each recognition pass, making the overall speech recognition process more efficient, faster and accurate.

BRIEF DESCRIPTION OF THE DRAWINGS

[0008] **Figure 1** is a block diagram of a speech recognition system that operates in accordance with the present invention;

[0009] **Figure 2** is a flow chart illustrating a method for recognizing words that have observable relationships; and

[0010] **Figure 3** is a flow chart illustrating a method for generating or selecting new language models and/or new acoustic models for use in a speech recognition process.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

[0011] Figure 1 is a block diagram illustrating a speech recognition system 101 that operates in accordance with the present invention. This system 101 may be implemented in a portable device such as a hand held computer, a portable phone, or an automobile. It may also be implemented in a stationary device such as a desktop personal computer or an appliance, or it may be distributed between both local and remote devices. The speech recognition system 101 illustratively comprises a speech recognition front end 103, a speech recognition engine 105, a processor 107, and a memory/database 109.

[0012] The speech recognition front end 103 receives and samples spoken input, and then measures and extracts features or characteristics of the spoken input that are used later in the speech recognition process. The speech recognition engine 105 may include a search algorithm (such as a Viterbi search algorithm) and acoustic models (such as models of individual phonemes or models of groups of phonemes) used in the speech recognition process. The processor 107 and associated memory 109 together operate as a computer to control the operation of the front end 103 and the speech recognition engine 105. The memory 109 stores recognizable words and word sets 111 in an accessible database that is used by the system 101 to process speech. Memory 109 also stores the software 115 that is used to implement the methods of the present invention. Both the speech recognition front end 103 and the speech recognition engine 105 may be implemented in hardware, software, or combination of hardware and software. All of the elements 103-109 may communicate with each other as required.

[0013] The invention relates to speech recognition systems and methods used to recognize words that have observable relationships. Examples of word sets with observable relationships are addresses; locations; names and telephone numbers; airline flight numbers, departure/arrival times, and departure/arrival cities; product part numbers, catalog numbers, and product names; and any other sets of words used to identify a person, place or thing.

[0014] Groups of words with observable relationships may be referred to as “sparse domains” or domains that have a small “Cartesian product” because typically only a small fraction of all possible word combinations are valid combinations. For example, an address with the ZIP code “94025” is only associated with the city of Menlo Park, California. “San Francisco, California 94025” or “Menlo Park, New Jersey 94025” are not valid addresses.

[0015] Figure 2 is a flow chart illustrating a preferred method for recognizing words that have observable relationships. This method may be implemented as a software routine 115 that is executed by the processor 107 of Figure 1. When a speech signal that represents a spoken utterance is received (step 201), a speech recognition “pass” is performed by applying a first language model to the speech signal (step 203). The language model may be a probabilistic finite state grammar, a statistical language model, or any other language model that is useful in a speech recognition system. The first recognition pass does not attempt to recognize the entire speech signal; for example, if the utterance represents an address, the first recognition pass may use a language model that recognizes only city names or only street numbers.

[0016] Next, a new language model and/or new acoustic models are selected or generated (step 205). The selection or generation of the new model or models is based at least in part on results from the previous recognition pass, and may also be based on information regarding the linguistic structure of the domain and/or information regarding relationships among concepts, objects, or components in the domain. For example, the previous recognition passes may have recognized the city name "Menlo Park" and the street number "333." Based on this information, a new language model might be generated or selected that includes only those streets in Menlo Park that have "333" as a street number.

[0017] This new language model and/or acoustic models and at least a portion of the speech signal are then used to perform another recognition pass (step 207). If a satisfactory recognition of the spoken utterance is complete (step 209), the speech recognition process ends (step 211). If a satisfactory recognition of the spoken utterance is not complete, then steps 205-209 are repeated as necessary.

[0018] Figure 3 is a flowchart that illustrates a preferred method for generating or selecting a new language model and/or new acoustic models (i.e., a method performing step 205 of FIG. 2.). In this method, a result from a speech recognition pass is acquired (step 301). This result includes a component, object or concept of the relevant domain. For example, if the speech recognition system is being used to recognize an address, the result from the previous recognition pass may include a street number or city name.

[0019] Next, the result from the speech recognition pass is used to perform a search on a database that contains information regarding relationships among the domain concepts, objects, or components (step 303). For example, the database may be a relational database that has information regarding the relationships among the components of an address. A search on the city name "Menlo Park" might find all the street names in that city; a search on the ZIP code "94025" might find all the streets within that ZIP code; and so on.

[0020] Finally, one or more results from the database search are then used to select or generate a language model and/or acoustic models (step 305). For example, the results from a database search on the ZIP code "94025" might be used to generate a language model (or select an existing language model) that includes all of the street names in that ZIP code. Or, the results from a database search on the city name "Menlo Park" and the street name "Ravenswood Avenue" might be used to generate or select a language model that includes all of the street numbers on Ravenswood Avenue in Menlo Park. Language models generated or selected this way can be used to greatly reduce the search space of subsequent recognition passes, making the speech recognition process both faster and more accurate.

[0021] While foregoing is directed to the preferred embodiment of the present invention, other and further embodiments of the invention may be devised without departing from the basic scope thereof, and the scope thereof is determined by the claims that follow.